

**UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK**

LYNNE FREEMAN,

Plaintiff,

v.

TRACY DEEBE-ELKENANEY P/K/A TRACY
WOLFF, et al.,

Defendants.

Case No. 1:22-cv-02435-LLS-SN

**DECLARATION OF CAROLE CHASKI
IN SUPPORT OF PLAINTIFF'S
MOTION TO EXCLUDE EXPERT
TESTIMONY**

DECLARATION OF CAROLE CHASKI

I, Carole Chaski, hereby declare as follows:

1. I am above eighteen (18) years of age and am competent to give the testimony set forth below. Said testimony is given from my own personal knowledge except where otherwise indicated. If called as a witness, I could and would competently testify as set forth below.
2. Linguistics is a broad field. Simply being a linguist does not make such linguist an expert in all types of linguistics. Further, linguistics is not well-known outside of academe. Therefore, it is easy to mislead a non-linguist about expertise in linguistics and the relevance of an expert's actual expertise. Professor Coulthard's actual expertise in discourse analysis is not used in his evaluation of my report.
3. To my knowledge, Professor Coulthard has never published a dataset for forensic linguistic evidence, nor has he ever independently collected ground-truth data for forensic

linguistic evidence.¹ This stands in sharp contrast to other forensic linguists who have collected and published datasets for forensic linguistics.

4. This is important because part of collecting data sets is curating them to ensure that they can do / test certain things. For example, in this case prior to undertaking my analysis I had to ensure that the baseline data was a fair sampling of works (e.g., in the same genre, timeframe, and quality). It was irrelevant to my ability to curate those works that I did not collect them. That said, understanding the computational and statistical analysis we are undertaking—which I do not believe Coulthard does—is critical to understanding whether the subject data set is proper.
5. Data collection is a highly valued skill in forensic linguistics, which, however, Professor Coulthard has never demonstrated. Below are examples a few dataset collection efforts with which I and my colleagues have been involved:
 - a. (1997) I collected the first dataset for forensic authorship attribution.² This data collection of ground truth data was funded by the US Department of Justice, National Institute of Justice.³
 - b. (2005) Schler, Koppel, Argamon and Pennebaker collected a corpus of blogs for forensic linguistic profiling. This blog was used in research on predicting age and

¹ Ground truth data is data for which the correct answer is independently verifiable; in Chaski's dataset, the author of each document was independently known. Thus, such data can be used to experimentally test methods and determine the error rate of the method. This kind of data collection is totally different from "case by case" data in which Professor Coulthard simply takes whatever data he is given for any case and assumes that the data is accurate.

² Chaski, Carole E. 1997. "Who Wrote It? Steps Toward a Science of Authorship Identification." NATIONAL INSTITUTE OF JUSTICE JOURNAL. September 1997. Also available through National Criminal Justice Reference Service: <http://www.ncjrs.org> NCJ 184604.

³ Chaski, Carole E. 2001. "Empirical Evaluation of Language-Based Author Identification Techniques." FORENSIC LINGUISTICS: INTERNATIONAL JOURNAL OF SPEECH, LANGUAGE AND LAW. Volume 8:1. pp. 1-64. June 2001. Published by University of Birmingham, England.

gender. The corpus was also available for download from Koppel's website for many years.⁴

- c. (2007) Forensic linguists at the FBI Behavioral Analysis Unit created the Communicative Threat Analysis Database, known as CTAD. This dataset has been used in at least two PhD dissertations.⁵
- d. (2009) Mihalcea and Strapparava collected a dataset of deceptive and truthful documents for deception detection research in forensic linguistics.⁶
- e. (2011) Chaski and Huddle collected the first dataset for forensic linguistic suicide note authentication.⁷ This dataset has been used in at least one PhD dissertation.
- f. (2019) Aston Institute for Forensic Linguistics initiated a project called FoLD, a repository of datasets. (Note that this project was initiated many years after Professor Coulthard retired from Aston University).⁸
- g. (2019) Almela, Alcaraz-Mármol, Garcia-Pinar and Pallejá collected writing samples from imprisoned domestic abusers and from university students at two universities in Spain.⁹ Almela and her colleagues used my Writing Sample Dataset

⁴ See for example all of the case studies that Coulthard reports in his textbook and articles. He publishes case studies but no experiments.

⁵ Fitzgerald, James R. 2007. "FBI's Communicated Threat Assessment Database: History, Design and Implementation." FBI LAW ENFORCEMENT BULLETIN, Volume 76 Issue 2. February 2007 pages 6-9. Also available through National Criminal Justice Reference Service: <http://www.ncjrs.org> NCJ 217407

⁶ Milhacea, R. and Strapparava, C. 2009. "The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language." PROCEEDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, ACL-IJCNLP 2009, pages 309-312, Singapore.

⁷ Chaski, Carole E. and Huddle, Denise. 2011. "Is This a Real Suicide Note? Authentication using Statistical Classifiers and Computational Linguistics." PROCEEDINGS OF THE ANNUAL MEETING OF THE AMERICAN ACADEMY OF FORENSIC SCIENCES.

⁸ <https://fold.aston.ac.uk/>

⁹ Almela, A., Alcaraz-Mármol, G., Garcia-Pinar, A. and Pallejá, C. 2019. "Developing and Analyzing a Spanish Corpus for Forensic Purposes." LINGUISTIC EVIDENCE IN SECURITY, LAW AND INTELLIGENCE. Volume 3, 1-13. Doi.org/10.5195/lesli.2019.19

prompts.

6. To my knowledge, Professor Coulthard has never published standards for data requirements for forensic linguistic evidence. Again, this stands in sharp contrast to other forensic linguists who have published guidelines for data collection and data requirements for forensic linguistics, including the following:

- a. (2005) I presented standard operating protocols for data quantity and statistical analysis for forensic authorship attribution.¹⁰
- b. (2013) I presented standard operating protocols for data quantity and statistical analysis for forensic authorship attribution.¹¹
- c. (2014) I presented guidelines and methods for collecting data from the Internet, at the American Academy of Forensic Sciences Annual Meeting.¹²
- d. (2014) I organized a symposium on forensic linguistic data at the Linguistic Society of America Annual Meeting. Symposia are competitively selected for inclusion in the annual meeting. This symposium included talks on the need for datasets,¹³ managing human subjects data according to Federal regulations,¹⁴ and

¹⁰ Chaski, Carole E. 2005. ““Who’s At the Keyboard? Recent Results in Authorship Attribution.” *INTERNATIONAL JOURNAL OF DIGITAL EVIDENCE*. Volume 4:1. Spring 2005. Available at <http://www.ijde.org>

¹¹ Chaski, Carole E. 2013. “Best Practices and Admissibility in Forensic Author Identification.” *JOURNAL OF LAW & POLICY*, Brooklyn Law School, Brooklyn, New York.

¹² Chaski, Carole E. 2014. “Collecting Ground-Truth, Web-Based Data for Research in Forensic Linguistics.” *PROCEEDINGS OF THE ANNUAL MEETING OF THE AMERICAN ACADEMY OF FORENSIC SCIENCES*.

¹³ Chaski, Carole E. 2014. “Data for Empirical Foundations in Forensic Linguistics: Overview of Symposium.” *SYMPOSIUM ON DATA FOR EMPIRICAL FOUNDATION IN FORENSIC LINGUISTICS*, LINGUISTIC SOCIETY OF AMERICA ANNUAL MEETING, Minneapolis, MN.

¹⁴ Parker, Judith A. and Chaski, Carole E. “Collecting Forensic Linguistic Data: Experimental Subjects and Authorship Identification.” *SYMPOSIUM ON EMPIRICAL FOUNDATIONS IN FORENSIC LINGUISTICS*, LINGUISTIC SOCIETY OF AMERICA ANNUAL MEETING, Minneapolis, MN. With Judith A Parker.

managing data from police and investigative agency sources.¹⁵

- e. (2015) Juola published a standard for using a distractor dataset as a baseline during authorship identification analysis.¹⁶
- 7. Coulthard has not demonstrated experience or expertise with data collection, data requirements or data standards. Instead, he studiously avoids providing any answers to questions about gold-standard data, data acquisition methods, data quality, data quantity, case versus baseline data, and any other issues about data collection. His textbook provides no answers to any questions about data collection.
- 8. Coulthard also lacks training and demonstrated experience in experimental linguistics, i.e. the design and conduct of experiments testing linguistic hypotheses. Computational linguistics includes learning about experimental design, design of stimuli, collection of experimental results and statistical analysis of results. Yet Coulthard has never published any experimental results in which he tested a method using ground truth data in order to provide an error rate. Instead, Professor Coulthard publishes case studies in which he reports how he “solved” a case, i.e. got the result that his client wanted, using whatever method he came up with for the case.
- 9. I am also unaware of Coulthard having any demonstrated expertise in computational linguistics. Text analysis by computer algorithm is a long-standing part of linguistics, with computational linguistics for machine translation starting in the 1950’s. Text analysis by

¹⁵ Barksdale, Larry (Sgt), Reddington, Michael, and Chaski, Carole E. 2014b. “Collecting Forensic Linguistic data: Police and Investigative Sources of Data for Deception Detection Research.” SYMPOSIUM ON DATA FOR EMPIRICAL FOUNDATION IN FORENSIC LINGUISTICS, LINGUISTIC SOCIETY OF AMERICA ANNUAL MEETING, Minneapolis, MN.

¹⁶ Juola, Patrick. 2015. “The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions.” DIGITAL SCHOLARSHIP IN THE HUMANITIES, Volume 30, Issue suppl_1, December 2015. Pages 100-113. Doi.org/10.1093/llc/fqv040

computer algorithm provides an objective, fatigue-resistant way to quantify patterns of language. Coulthard has published statements that he supports the use of computational tools in forensic linguistics, but to my knowledge he has never built any.

10. Finally, I am unaware of Coulthard having any demonstrated expertise in either statistical analysis or any machine learning techniques.

11. On this last point, I note that in the *Yukos* case which Coulthard mentions in his report, he partnered with his colleague Tim Grant who was responsible for responding to the statistical analysis. Attached hereto as Exhibit "1" is a true and correct copy of that report, which provides Professor Grant's qualifications as an expert in statistical analysis but includes no such qualifications for Coulthard. And, in any event, I showed in my counter-rebuttal in that case that their statistical critiques were not only wrong, but contradicted by the main statistics textbook that Professor Grant relied on, as well as other sources in statistics and machine learning that I referenced.

12. As noted above, the datasets I used in this case were specifically vetted to ensure that they were appropriate for the computational and statistical analyses I conducted. Because Coulthard has no training and/or demonstrated experience employing such methodologies he has no competent basis to tell this Court (or a jury) that my datasets are unreliable.

13. In addition to being an expert in syntax and language change, I am expert in computational linguistics. *Id.* at ¶19. Computational linguistics is a combination of computer science and linguistics which is the technology underlying Google search, Microsoft Word and current developments such as ChatGPT, QuillBot paraphrasing tool, Spinbot for text rewriting, and other text spinning software. *Id.* As a computational linguist, I have worked on teams that developed spell checkers, grammar checkers and question-answering systems and I have developed software for forensic text analysis, coining the terms

forensic computational linguistics and computational forensic linguistics to describe the work that I and now others do. Id. I have taught graduate and undergraduate courses in computational linguistics. Id. at 20.

14. In order to fairly evaluate my report, an expert would need to have demonstrated expertise in four areas: (i) linguistic data collection; (ii) experimental linguistics; (iii) computational linguistics text analysis; and (iv) statistics or machine learning. Any doctoral-level individual who is trained in psycholinguistics, computational linguistics or computer science with a specialty in natural language processing and who has experience in linguistic standards for data collection is qualified in these four areas of normal science that I have applied in the context of forensic linguistic evidence. While there are quite a few experts who have this demonstrated expertise (e.g., Professor Pascual Cantos Gómez, Professor Angela Almela, Professor Efstathios Stamatatos, and Dr Blake Howald), Coulthard does not.

I declare under penalty of perjury and the laws of the United States of America that the foregoing is true and correct.

Executed on December 22, 2023 in Los Angeles, California.

By: Carole E. Chaski PhD
Carole E. Chaski, PhD
Declarant